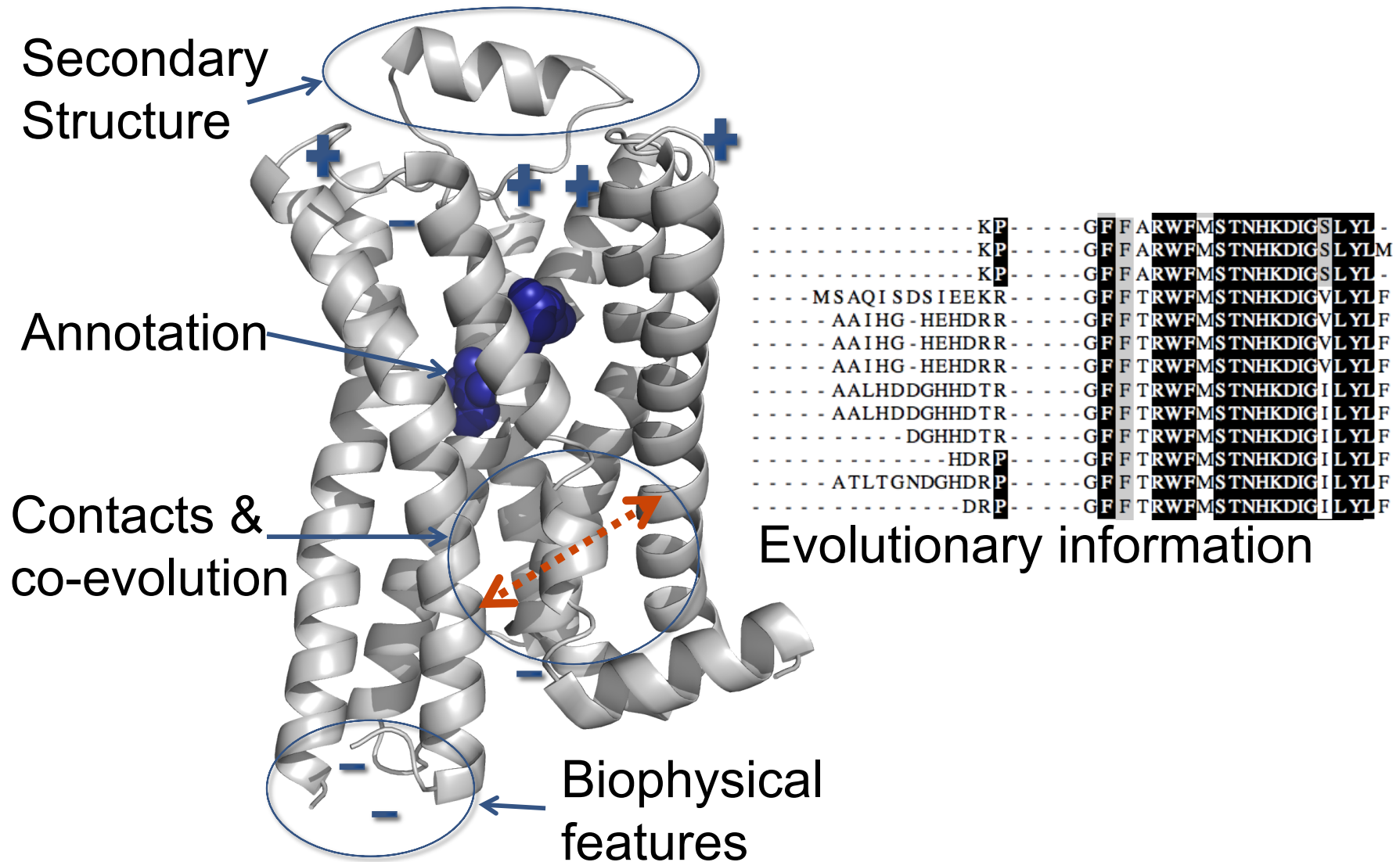


TUMseq meeting

SNAP

- Screening for Non-Acceptable Polymorphisms
- Predicts functional effects single AA-substitutions from sequence
- Neural Network-based classifier (i.e. Machine Learning)
- Originally developed by Yana Bromberg (2007)
- Winner of CAFA competition (2010)
- Re-trained and improved (2012)

How do we predict effect?

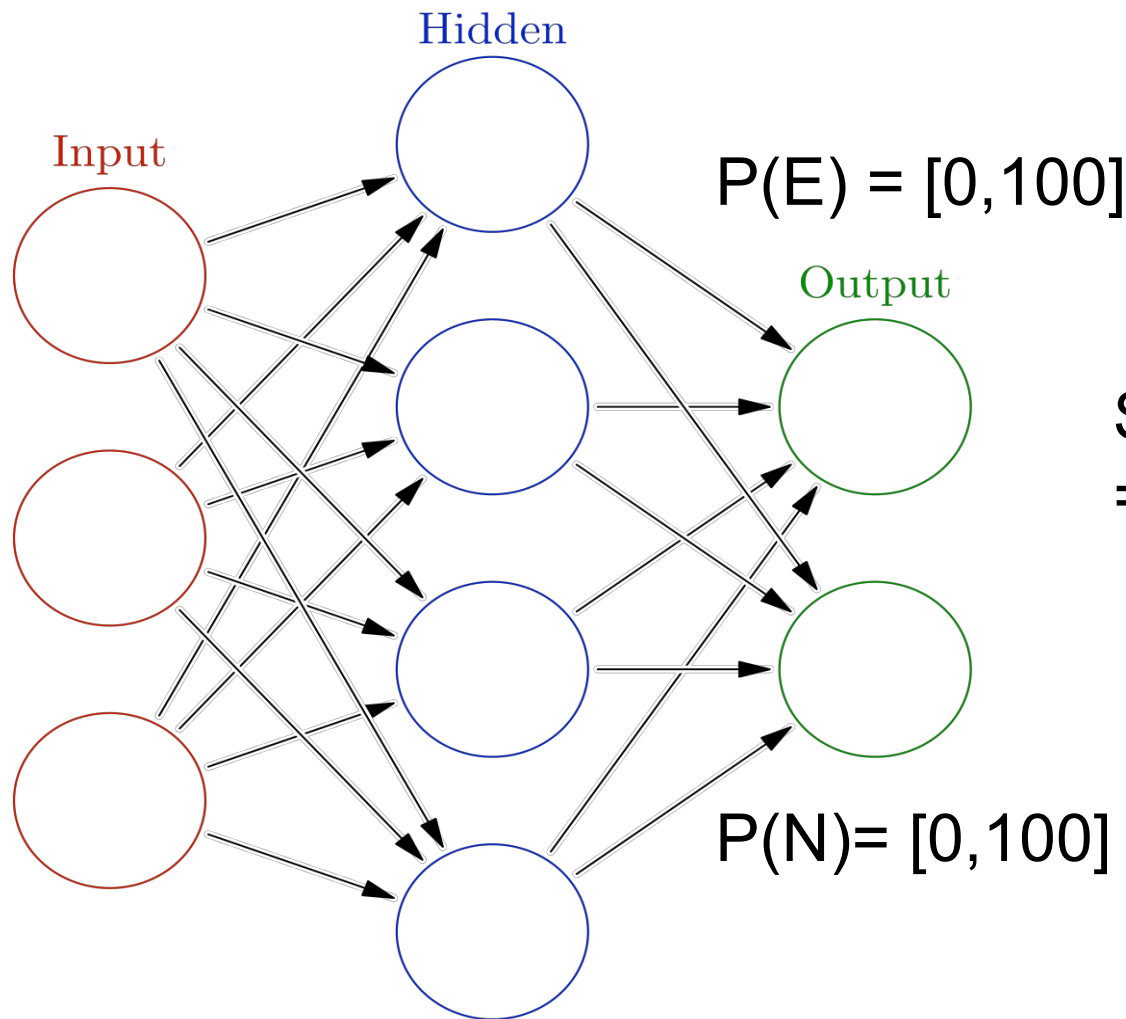


Our training data

- Trained on ~100k variants with annotated effects (supervised learning)
- Data sources:
 - PMD (Literature reports on functional effects)
 - OMIM and HUMVAR (Disease-related variants)
 - EC (Pseudo-neutral variants)

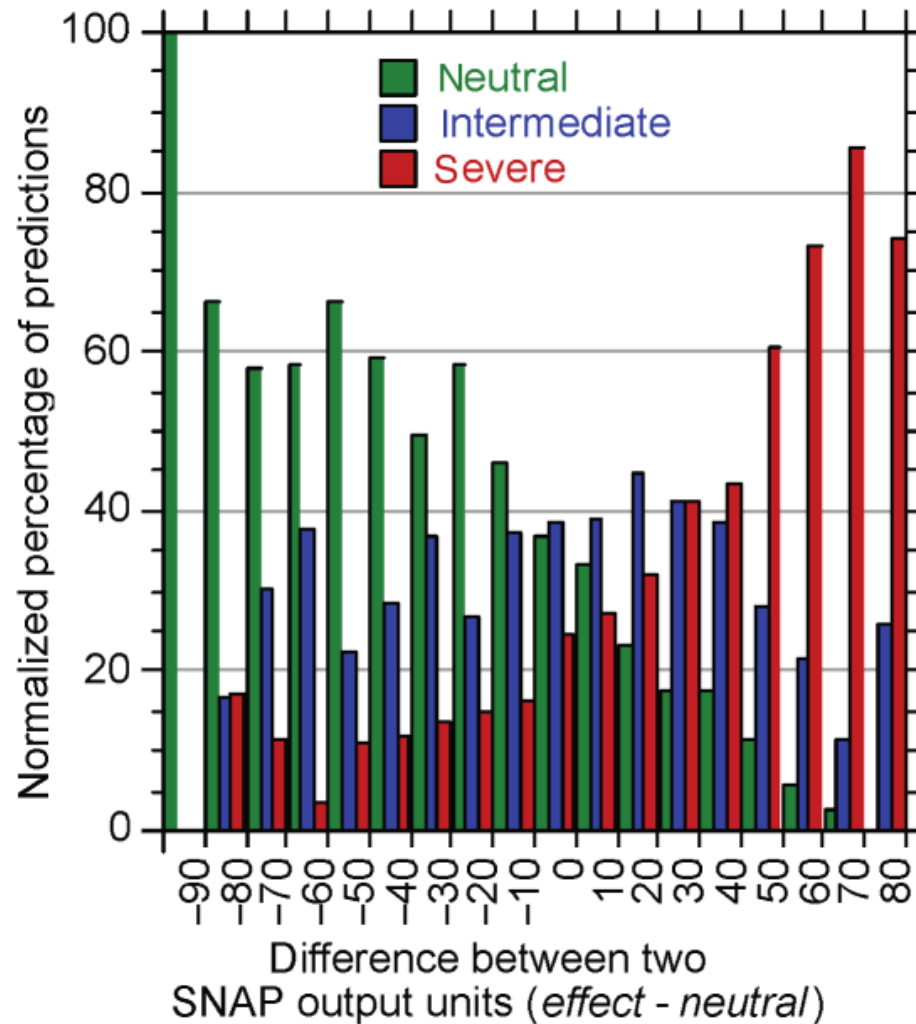
Source	Effect	Neutral	Total
PMD	38k	13k	51k
OMIM	23k	-	23k
EC	-	27k	27k
Total	61k	40k	101k

SNAP - Neural network with 2 outputs



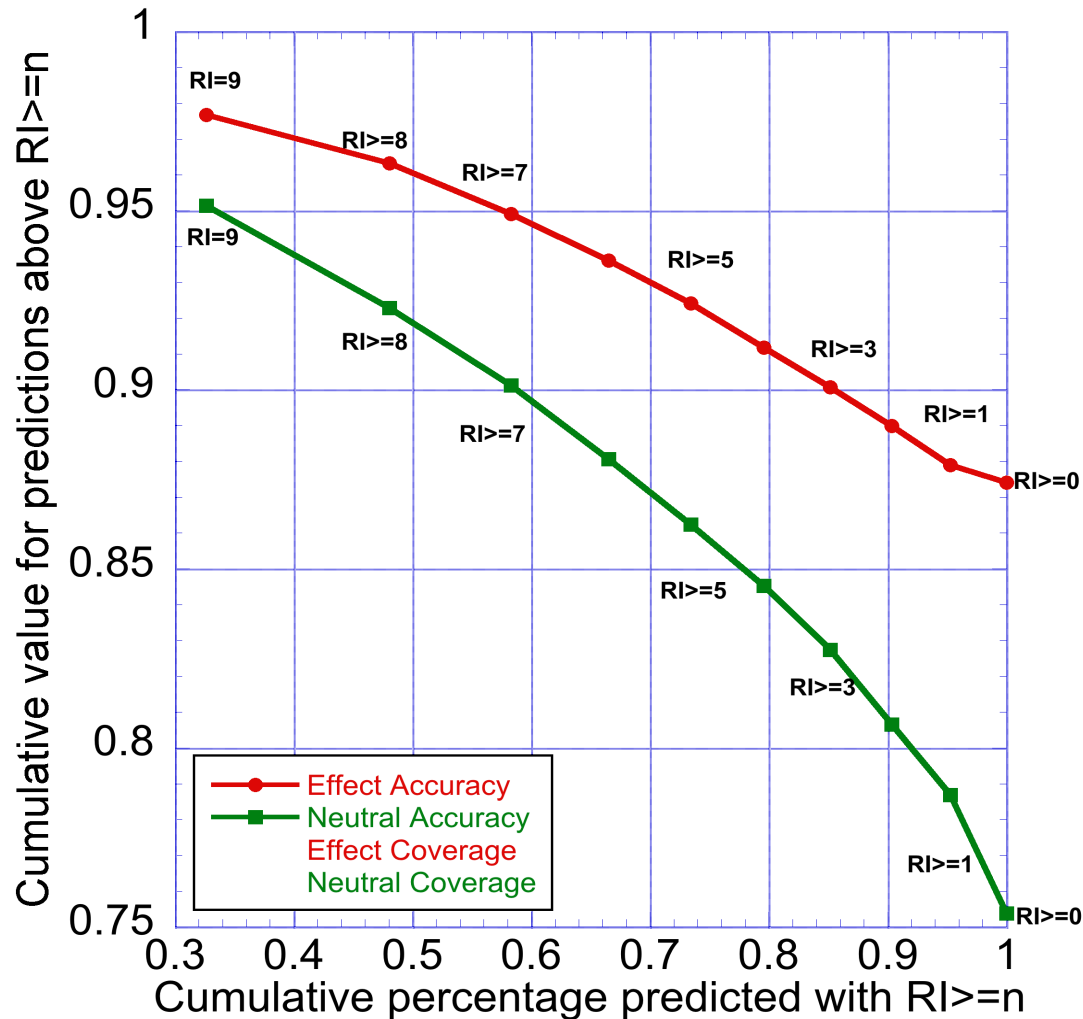
$$\text{SNAP Score} = P(E) - P(N) \\ = [-100, +100]$$

SNAP predicts severity of effect



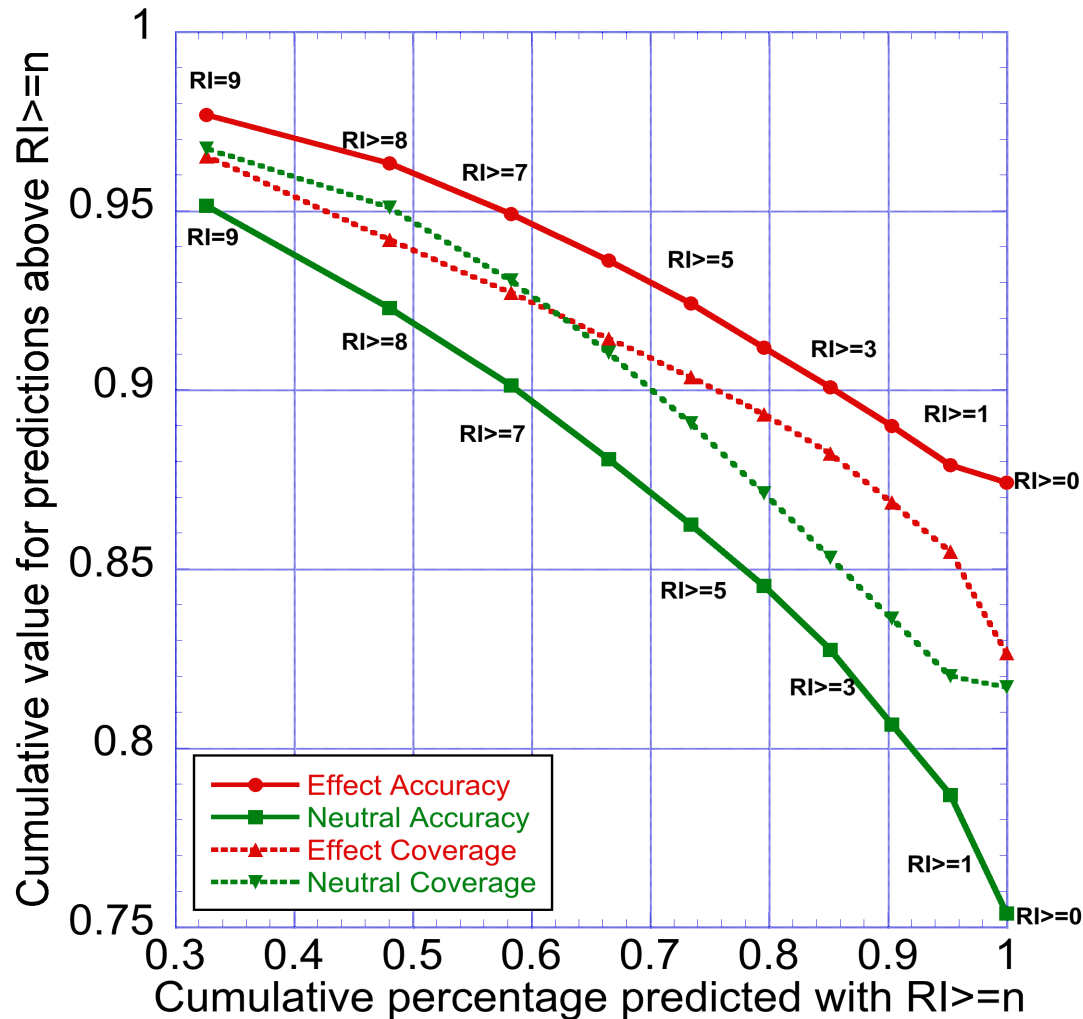
- SNAP score correlates with severity of effect
- Severe cases have high scores (>50)
- Neutral cases have low scores (<-50)
- Intermediates evenly distributed

RI indicates prediction confidence



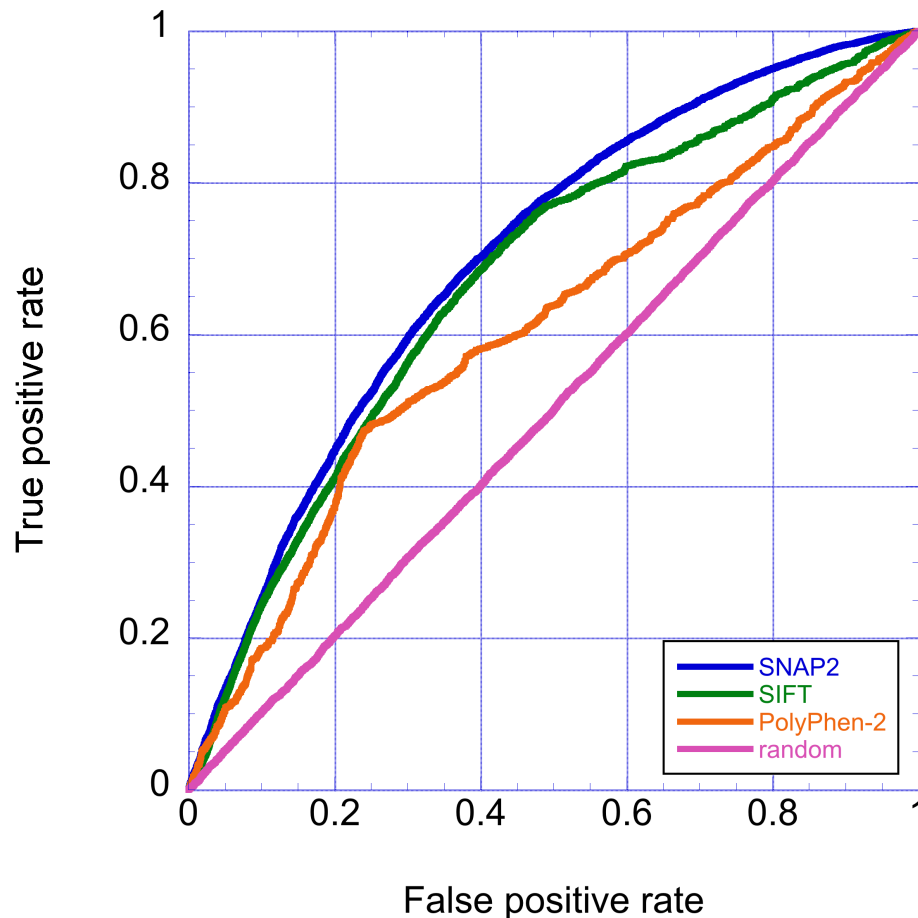
- RI calculated as difference:
 $p(\text{neutral}) - p(\text{effect})$
- Almost 60% predicted with $RI \geq 7$
- Accuracy at $RI \geq 7$:
Neutral: 90%
Effect: 95%

RI indicates prediction confidence



- RI calculated as difference:
 $p(\text{neutral}) - p(\text{effect})$
- Almost 60% predicted with $RI \geq 7$
- Accuracy at $RI \geq 7$:
Neutral: 90%
Effect: 95%

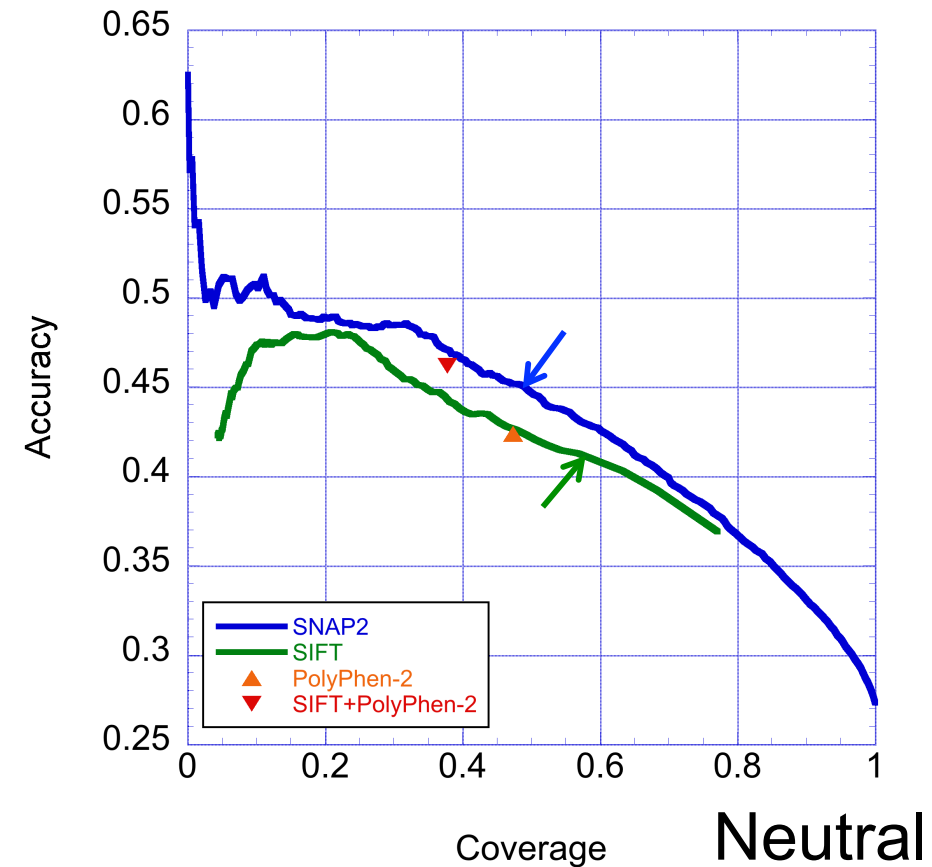
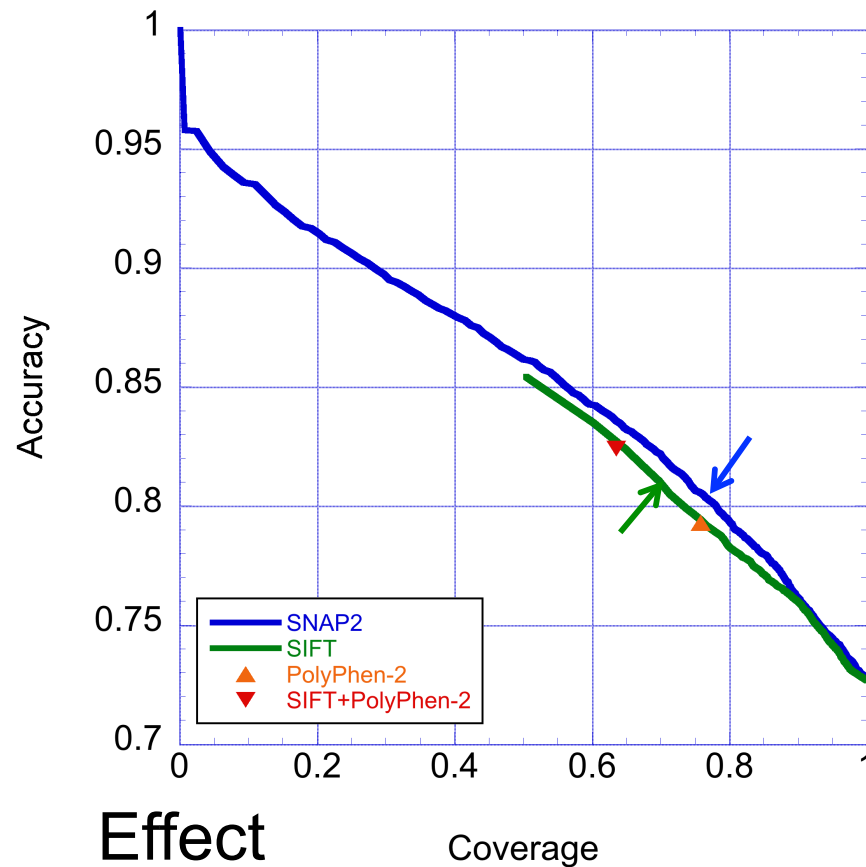
SNAP better than competitors



- Receiver operating characteristic (ROC)
- AUC:
 - SNAP: 0.70
 - SIFT: 0.67
 - PolyPhen-2: 0.62
 - Random: 0.50
- Tested on PMD set

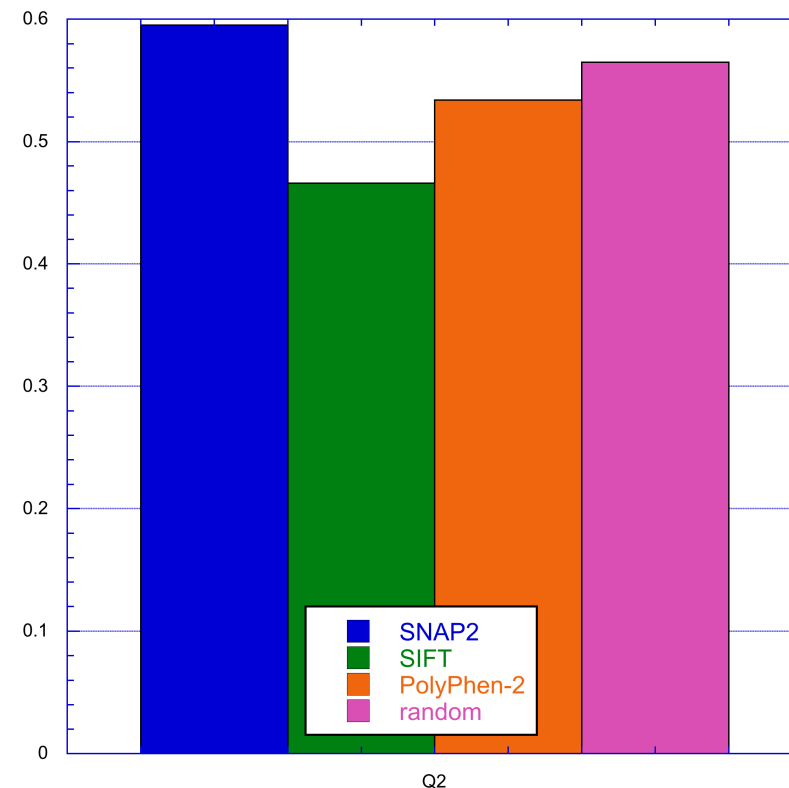
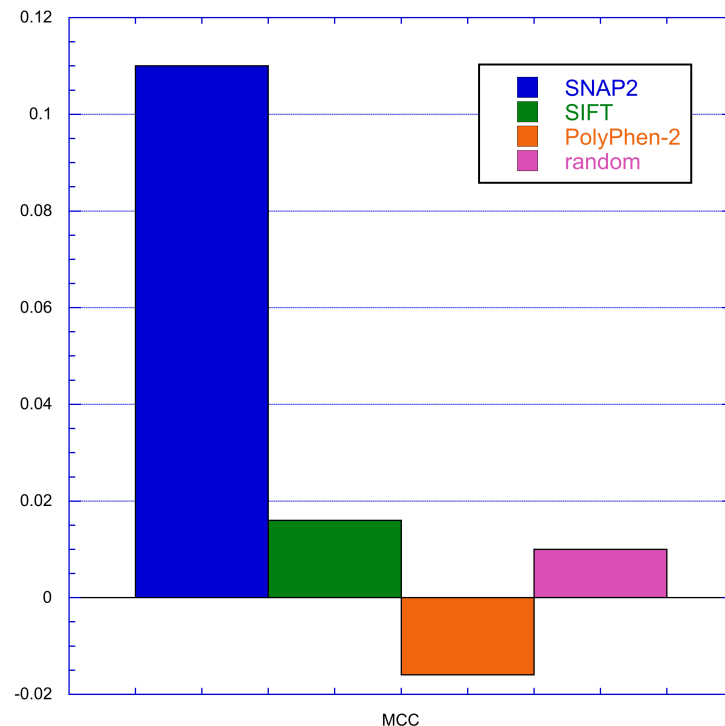
SNAP better than majority vote

- Tested on 28971 variants from PMD set



Prediction of “hard cases”

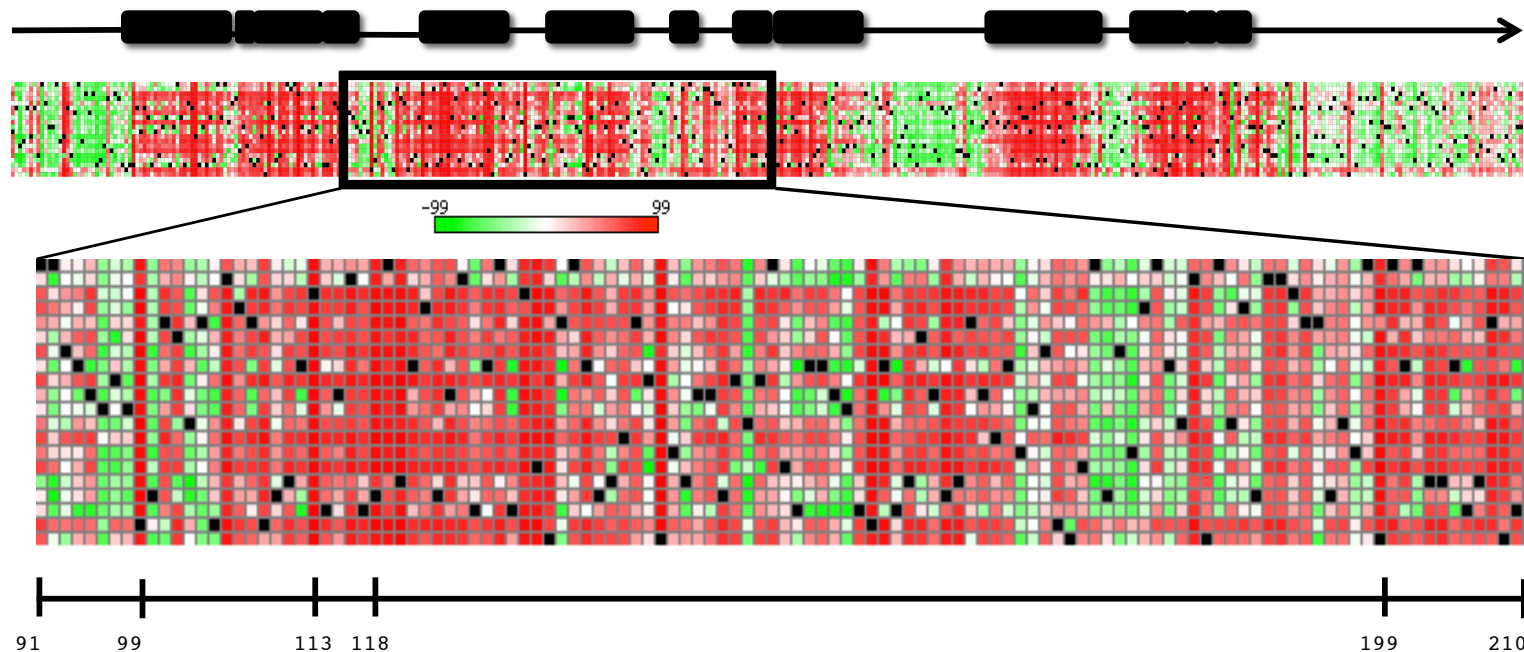
- SIFT and PolyPhen-2 disagree: 6267 (4k + /2k -)
- “smart” random assumes 65:35 background



$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

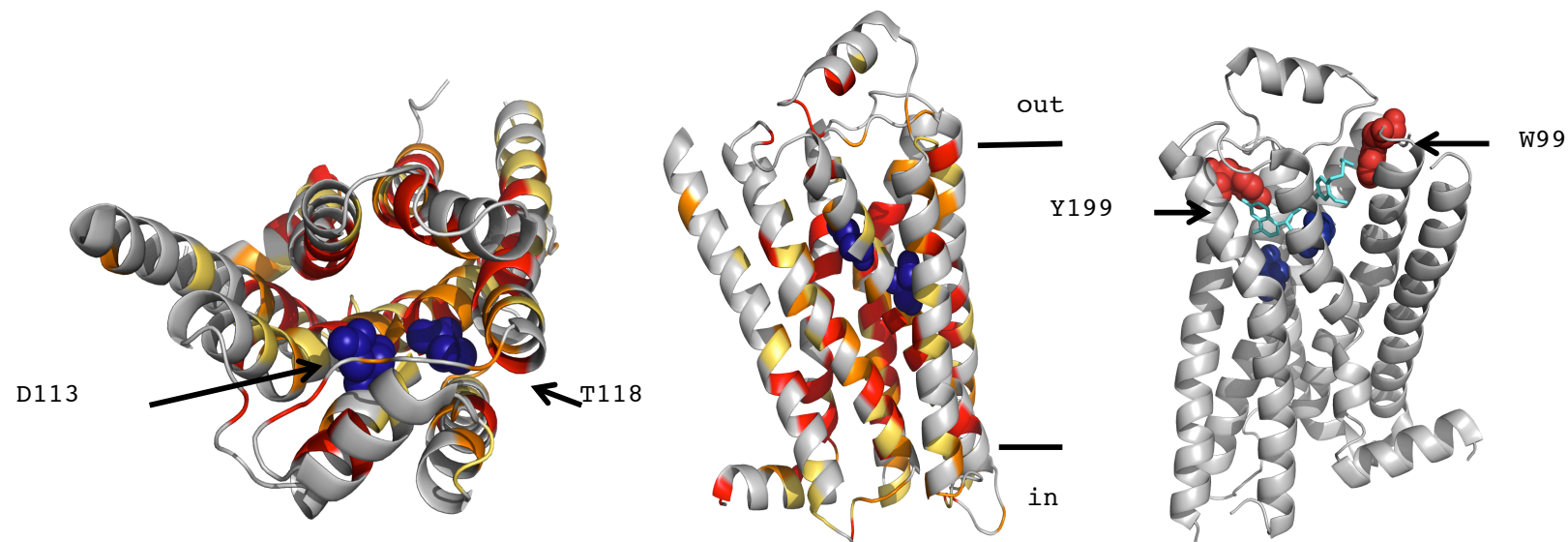
The mutability landscape

- 19-non-native: complete in-silico mutagenesis prediction
- Mutate every amino acid at each position into each other
- E.g.: Human beta-2 adrenergic receptor



The mutability landscape

- Apply scoring filters for structurally neutral variants and evolutionary constraints

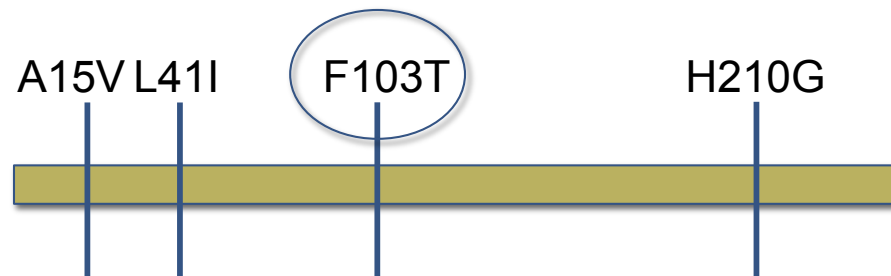


Detecting high risk SNPs

- Filter disease SNPs from natural variation
- Problem: Can we find the bad nsSNPs among all those present in an individual's genome?
- Data:
 - OMIM: known disease-causing SNPs in human
 - 1KG: naturally occurring SNPs in 1000 individuals

Genes / Mutations	Amount
Genes with annotated OMIM mutation	1549
OMIM Mutations on these genes	5050
1KG Mutations on these genes	27303
OMIM Mutations present in 1KG data	700

Ranking of OMIM Mutations



45 30

62

88

Determine SNAP Score

88, 62, 45, 30

Sort according to SNAP Score

1, 2, 3, 4

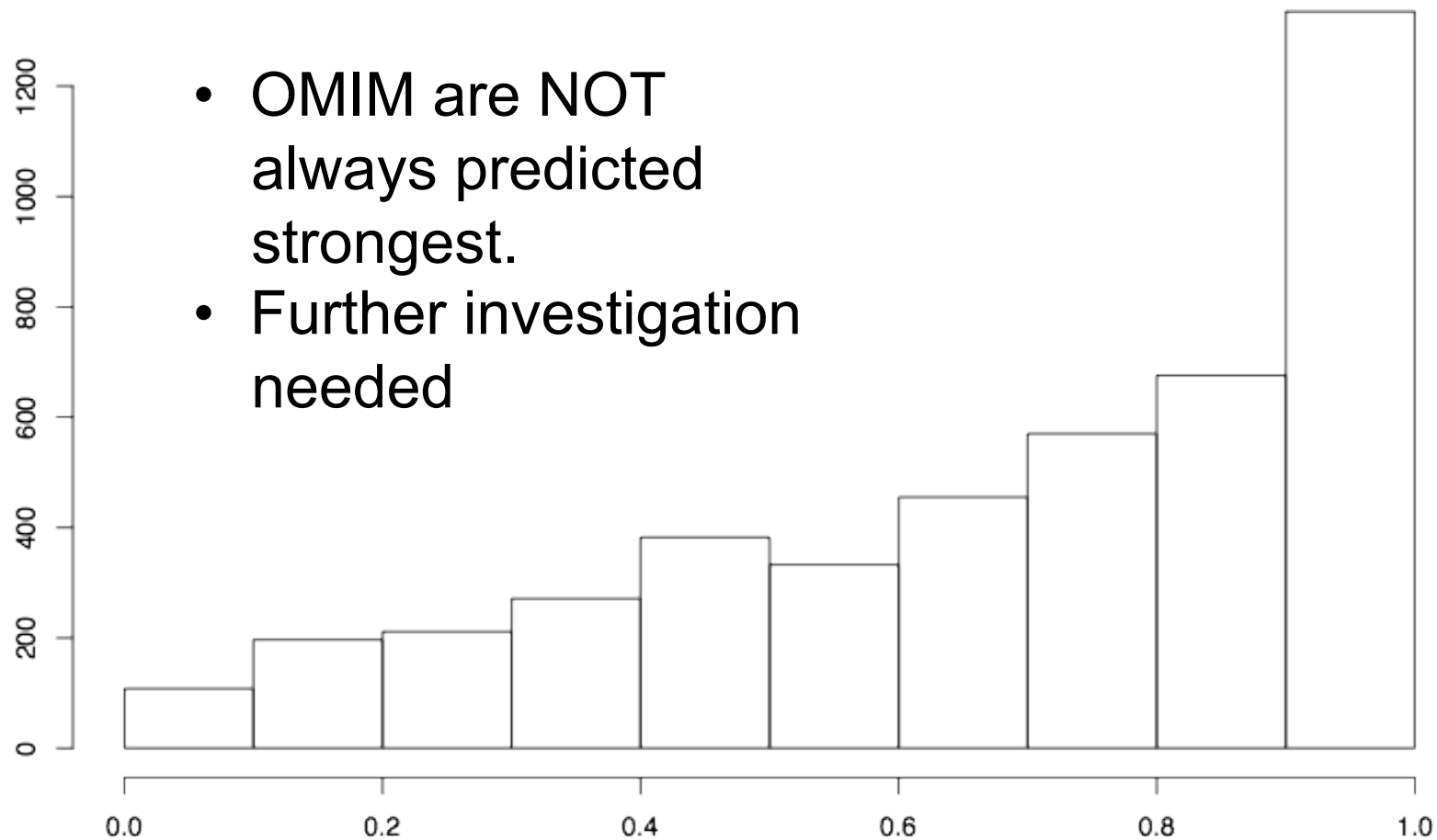
Assign Rank

1, 0,66, 0,33, 0

Normalize Rank to [0;1]

Our results so far:

- OMIM are NOT always predicted strongest.
- Further investigation needed



Possible contribution to TUMseq

- Optimize prioritization performance
- Train SNAP on DNA data
- Include motif/pattern detection (e.g. transcription factor binding) for non-coding SNPs
- Effects of synonymous SNPs
- Analyze Bull data – build method optimized on the data